W81

PATENT ABSTRACTS OF JAPAN

(11)Publication number : 11-259487

(43)Date of publication of application : 24.09.1999

(51)Int.Cl. G06F 17/30

(21)Application number : 10-055560

(22)Date of filing : 06.03.1998

(71)Applicant : TOSHIBA CORP
TOSHIBA COMPUT ENG CORP

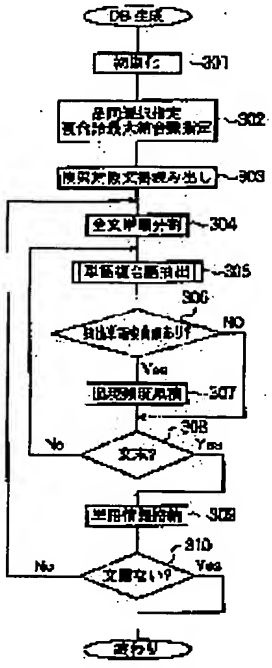
(72)Inventor : TANOSAKI YASUO
NISHINA TAKUYA
NAKAMOTO YUKIO
KUBOTA NAOHIDE

(54) SIMILAR DOCUMENT RETRIEVING DEVICE, SIMILAR DOCUMENT RETRIEVING METHOD AND STORAGE MEDIUM RECORDED WITH PROGRAM FOR RETRIEVING SIMILAR DOCUMENT

(57)Abstract:

PROBLEM TO BE SOLVED: To enhance retrieving accuracy when a similar document is retrieved by extracting a composite word from a retrieving key document and a document to be retrieved.

SOLUTION: When the composite word, i.e., 'brush character address printing function' exists in the retrieving key document or the document to be retrieved and the maximum number of connections is specified as three, all the composite words like 'brush character address', 'character address printing', 'address printing function', 'brush character', 'character address', etc., consisting of the number of words to be equal to or less than the maximum number of connections are extracted from each document, appearance frequency of the composite words is calculated and a degree of similarity between the retrieving key document and the document to be retrieved is calculated. Since different composite words to characterize the documents with specified contents are thoroughly extracted, a more proper degree of similarity between the documents is calculated and highly accurate retrieval of similar document intended by a user is performed.



LEGAL STATUS

[Date of request for examination]

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]
[Date of registration]
[Number of appeal against examiner's decision of rejection]
[Date of requesting appeal against examiner's decision of rejection]
[Date of extinction of right]

W81

(19)日本国特許庁（ＪＰ）　(12)公開特許公報（Ａ）　(11)特許出願公開番号

特開平11－259487

(43)公開日　平成11年(1999)９月24日

(51)Int.Cl.[6]	識別記号	FI		
Ｇ０６Ｆ　17/30		Ｇ０６Ｆ	15/403	３３０Ｂ
			15/401	３１０Ａ
			15/403	３５０Ｃ

審査請求　未請求　請求項の数５　ＯＬ　（全 12 頁）

(21)出願番号　特願平10－55560

(22)出願日　平成10年(1998)３月６日

(71)出願人　000003078
株式会社東芝
神奈川県川崎市幸区堀川町72番地

(71)出願人　000221052
東芝コンピュータエンジニアリング株式会社
東京都青梅市新町３丁目３番地の１

(72)発明者　田野崎　康雄
東京都青梅市末広町２丁目９番地　株式会社東芝青梅工場内

(74)代理人　弁理士　須山　佐一

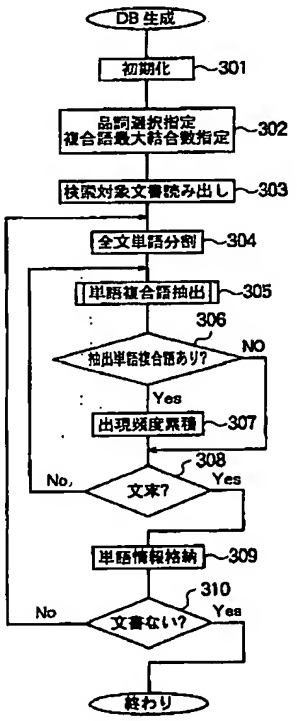
最終頁に続く

(54)【発明の名称】　類似文書検索装置、類似文書検索方法、および類似文書検索のためのプログラムが記録された記録媒体

(57)【要約】

【課題】　検索キー文書と検索対象文書から複合語を抽出して類似文書を検索する場合の検索精度の向上を図る。

【解決手段】　検索キー文書或いは検索対象文書に例えば「筆文字宛名印刷機能」といった複合語が存在し、最大結合数を３と指定したとき「筆文字宛名」「文字宛名印刷」「宛名印刷機能」「筆文字」「文字宛名」などの最大結合数以下の単語数からなる複合語を各文書からすべて抽出し、これらの複合語の出現頻度を計算して、検索キー文書と検索対象文書との類似度を算出する。特定の内容の文書を特徴付ける異なる複合語を漏れなく抽出することができるので、文書間のより妥当な類似度を計算でき、ユーザの意図する精度の高い類似文書検索を行うことができる。



【特許請求の範囲】
【請求項１】　ある文書を検索キー文書としてこの検索キー文書と類似する文書を複数の検索対象文書の中から検索する類似文書検索装置において、
前記検索キー文書および前記検索対象文書を単語単位に分割する分割手段と、
前記分割手段によって分割された単語の中から予め指定された条件を満たす単語を抽出する単語抽出手段と、
前記検索キー文書および前記検索対象文書から抽出すべき複合語を構成する単語数の上限値を指定する指定手段と、
前記単語抽出手段によって抽出された単語の結合により構成される複合語のうち前記指定手段により指定された上限値以下の数の単語により構成されるすべての複合語を前記検索キー文書および前記検索対象文書から抽出する複合語抽出手段と、
前記複合語抽出手段によって抽出された複合語の前記検索キー文書および前記検索対象文書での出現頻度をそれぞれ算出する手段とを具備することを特徴とする類似文書検索装置。
【請求項２】　ある文書を検索キー文書としてこの検索キー文書と類似する文書を複数の検索対象文書の中から検索する類似文書検索装置において、
前記検索キー文書および前記検索対象文書を単語単位に分割する分割手段と、
前記分割手段によって分割された単語の中から予め指定された条件を満たす単語を抽出する単語抽出手段と、
前記検索キー文書および前記検索対象文書から抽出すべき複合語を構成する単語の数の上限値を指定する指定手段と、
前記単語抽出手段によって抽出された単語の結合により構成される複合語のうち前記指定手段により指定された上限値以下の数の単語により構成されるすべての複合語を前記検索キー文書および前記検索対象文書から抽出する複合語抽出手段と、
任意の単語を不要語として選択する不要語選択手段と、
前記複合語抽出手段によって抽出された複合語のうち前記不要語選択手段によって選択された不要語を含む複合語を無効とする複合語無効化手段と、
前記複合語抽出手段によって抽出された有効な複合語の前記検索キー文書および前記検索対象文書での出現頻度をそれぞれ算出する手段とを具備することを特徴とする類似文書検索装置。
【請求項３】　請求項１または２記載の類似文書検索装置において、
前記単語抽出手段による単語の抽出条件として単語の品詞を指定する手段を有することを特徴とする類似文書検索装置。
【請求項４】　ある文書を検索キー文書としてこの検索キー文書と類似する文書を複数の検索対象文書の中から検索する類似文書検索方法において、
前記検索キー文書および前記検索対象文書を単語単位に分割し、
前記分割された単語の中から予め指定された条件を満たす単語を抽出し、
前記抽出された単語の結合により構成される複合語のうち予め指定された数以下の単語で構成されるすべての複合語を前記検索キー文書および前記検索対象文書から抽出し、
前記抽出された複合語の前記検索キー文書および前記検索対象文書での出現頻度を算出することを特徴とする類似文書検索方法。
【請求項５】　ある文書を検索キー文書としてこの検索キー文書と類似する文書を複数の検索対象文書の中から検索するためのプログラムが記録された記録媒体であって、
前記検索キー文書および前記検索対象文書を単語単位に分割する分割手段と、
前記分割手段によって分割された単語の中から予め指定された条件を満たす単語を抽出する単語抽出手段と、
前記検索キー文書および前記検索対象文書から抽出すべき複合語を構成する単語数の上限値を指定する指定手段と、
前記単語抽出手段によって抽出された単語の結合により構成される複合語のうち前記指定手段により指定された上限値以下の単語数で構成されるすべての複合語を前記検索キー文書および前記検索対象文書から抽出する複合語抽出手段と、
前記複合語抽出手段によって抽出された複合語の前記検索キー文書および前記検索対象文書での出現頻度をそれぞれ算出する手段とを具備するプログラムが記録されていることを特徴とする記録媒体。
【発明の詳細な説明】
【０００１】
【発明の属する技術分野】本発明は、文書データベースから、文書間の類似度に基づく文書データの検索を行う類似文書検索装置、類似文書検索方法、および類似文書検索のためのプログラムが記録された記録媒体に関する。
【０００２】
【従来の技術】近年、大量の電子化された文書データが流通するようになり、自動分類等を行う目的で、文書データベース中から指定された文書（以下、検索キー文書と呼ぶ）に類似する文書の自動検索を行うシステムが実用化されてきている。この文書検索システムでは、検索キー文書に含まれている単語と検索対象となる文書（以下、検索対象文書と呼ぶ）に含まれている単語とを比較し、共通する単語の種類、出現場所、出現回数などから空間ベクトル法により類似度を算出して、類似度の高い検索対象文書を検索結果として出力する。

【０００３】このような類似文書検索では、検索キー文書や検索対象文書から、その文書を特徴付ける単語を抽出することが、精度の高い類似文書検索を行うために非常に重要な鍵となる。これまで、文書から名詞やサ変名詞などの単語を対象に単語の抽出を行っていたが、文書から抽出された単一の単語が必ずしもその文書を特徴付ける単語として使用されているとは限らない。

【０００４】そこで、文書を特徴付ける複数の単語の結合からなる複合語を文書から抽出する方法が提案されている。このように文書を特徴付ける複合語を抽出することで、単位の単語を抽出する方法に比べ、ユーザの意図する検索結果をより高い精度で得ることができる。

【０００５】

【発明が解決しようとする課題】ところで、このように複合語を抽出する方法では、複合語の語長（単語数）が長くなればなるほど、その複合語がその文書を特徴付ける度合も高くなるが、その反面、検索対象の文書から抽出される単語の数が極端に減ってしまう傾向がある。このため、本来類似文書として検索されるべき文書に対する類似度として妥当な値が得られず、やはりユーザの意図する検索結果を高精度に得ることは困難であった。

【０００６】本発明は、このような課題を解決するためのもので、検索キー文書と検索対象文書から複合語を抽出して類似文書を検索する場合の検索精度の向上を図ることのできる類似文書検索装置、類似文書検索方法、および類似文書検索のためのプログラムが記録された記録媒体の提供を目的とする。

【０００７】

【課題を解決するための手段】上記した目的を達成するために、本発明は、請求項１に記載されるように、ある文書を検索キー文書としてこの検索キー文書と類似する文書を複数の検索対象文書の中から検索する類似文書検索装置において、前記検索キー文書および前記検索対象文書を単語単位に分割する分割手段と、前記分割手段によって分割された単語の中から予め指定された条件を満たす単語を抽出する単語抽出手段と、前記検索キー文書および前記検索対象文書から抽出すべき複合語を構成する単語数の上限値を指定する指定手段と、前記単語抽出手段によって抽出された単語の結合により構成される複合語のうち前記指定手段により指定された上限値以下の数の単語により構成されるすべての複合語を前記検索キー文書および前記検索対象文書から抽出する複合語抽出手段と、前記複合語抽出手段によって抽出された複合語の前記検索キー文書および前記検索対象文書での出現頻度をそれぞれ算出する手段とを具備することを特徴とする。

【０００８】本発明の類似文書検索装置では、文書から抽出された単語の結合により構成される複合語のうち、指定された上限値以下の数の単語により構成されるすべての複合語を検索キー文書および検索対象文書から抽出することで、文書を特徴付ける複合語を漏れなく抽出することができ、検索キー文書と検索対象文書との類似度をより高精度に算出することができる。

【０００９】また、本発明は、請求項２に記載されるように、ある文書を検索キー文書としてこの検索キー文書と類似する文書を複数の検索対象文書の中から検索する類似文書検索装置において、前記検索キー文書および前記検索対象文書を単語単位に分割する分割手段と、前記分割手段によって分割された単語の中から予め指定された条件を満たす単語を抽出する単語抽出手段と、前記検索キー文書および前記検索対象文書から抽出すべき複合語を構成する単語の数の上限値を指定する指定手段と、前記単語抽出手段によって抽出された単語の結合により構成される複合語のうち前記指定手段により指定された上限値以下の数の単語により構成されるすべての複合語を前記検索キー文書および前記検索対象文書から抽出する複合語抽出手段と、任意の単語を不要語として選択する不要語選択手段と、前記複合語抽出手段によって抽出された複合語のうち前記不要語選択手段によって選択された不要語を含む複合語を無効とする複合語無効化手段と、前記複合語抽出手段によって抽出された有効な複合語の前記検索キー文書および前記検索対象文書での出現頻度をそれぞれ算出する手段とを具備することを特徴とする。

【００１０】本発明の類似文書検索装置では、文書を特徴付ける複合語を漏れなく抽出することができ、検索キー文書と検索対象文書との類似度をより高精度に算出することができるとともに、予め指定された不要語の単語を含む複合語を無効なものとすることで、ユーザの意図をさらに反映したより一層高精度な類似文書検索を行うことが可能になる。

【００１１】

【発明の実施の形態】以下、図面を参照して本発明の実施形態を説明する。

【００１２】図１は本発明の実施形態である類似文書検索装置のハードウェア構成を示すブロック図である。

【００１３】同図に示すように、この類似文書検索装置は、キーボードなどの入力装置１、ＣＰＵおよびメモリなどから構成される制御装置２、類似文書の検索結果などを表示する表示装置３、および、文書データや類似文書検索のための各文書の単語情報や品詞辞書、不要語辞書などを格納する外部記憶装置４などから構成されている。外部記憶装置４に格納された品詞辞書、不要語辞書の構成を図６、図７にそれぞれ示す。

【００１４】図２にこの類似文書検索装置における制御装置１の構成を示す。制御装置１は、制御部１００とメモリ部２００から構成される。

【００１５】制御部１００は、初期化部１０１、入力部１０２、出力部１０３、抽出対象品詞設定部１０４、最大結合数設定部１０５、検索対象文書読み出し部１０

６、検索対象単語切り出し部１０７、検索対象単語複合語抽出部１０８、検索対象単語出現頻度算出部１０９、検索対象単語情報書き込み部１１０、検索キー文書入力部１１１、検索キー単語切り出し部１１２、検索キー単語複合語抽出部１１３、検索キー単語出現頻度算出部１１４、検索対象単語情報読み出し部１１５、共通単語抽出部１１６、類似度算出部１１７、検索結果出力部１１８などから構成される。メモリ部２００は、品詞情報バッファ部２０１、選択品詞情報バッファ部２０２、不要語情報バッファ部２０３、最大結合数バッファ部２０４、検索対象文書格納バッファ部２０５、検索対象全文分割単語格納バッファ部２０６、検索対象複合語候補格納バッファ部２０７、検索対象抽出複合語格納バッファ部２０８、検索対象単語情報格納バッファ部２０９、検索キー文書格納バッファ部２１０、検索キー全文分割単語格納バッファ部２１１、検索キー複合語候補格納バッファ部２１２、検索キー抽出複合語格納バッファ部２１３、検索キー単語情報格納バッファ部２１４、共通単語情報格納バッファ部２１５、算出類似度格納バッファ部２１６、検索結果出力バッファ部２１７などから構成される。

【００１６】初期化部１０１は、上記各バッファ部の初期化を行い、更に、外部記憶装置４における辞書（品詞辞書、不要語辞書など）の内容をメモリ部に読み込む。

【００１７】入力部１０２は、ユーザによる入力装置１からの検索キー文書や単語抽出条件の設定など各種設定の入力を行う。

【００１８】出力部１０３は、入力部１０２により入力された検索キー文書などの各種設定内容を表示装置３に出力する。

【００１９】抽出対象品詞設定部１０４は、ユーザが品詞情報バッファ部２０１から選択した抽出対象単語の品詞を選択品詞情報バッファ部２０２に格納する。

【００２０】最大結合数設定部１０４は、ユーザが指定した複合語の最大結合数を最大結合数バッファ部２０４に格納する。

【００２１】検索対象文書読み出し部１０６は、外部記憶装置４に格納されている検索対象文書に関する情報を文書データベース化するために、文書データベース化すべき文書情報を外部記憶装置４から読み込み、検索対象文書格納バッファ部２０５に格納する。

【００２２】検索対象単語切り出し部１０７は、検索対象文書格納バッファ部２０５に格納されている検索対象文書からの単語切り出しを行う。そして、その検索対象文書から抽出される全ての単語とその品詞を検索対象文書全文分割単語格納バッファ部２０６に格納する。単語の切り出しは形態素解析などにより行い、その文書から抽出される単語の品詞情報を「名詞」、「サ変名詞」、「形容詞」などで表現する。

【００２３】検索対象単語複合語抽出部１０８は、検索対象全文分割単語格納バッファ部２０６に格納されている全ての単語とその品詞情報の中から、選択品詞情報バッファ部２０２に格納されている品詞情報を参照して、該当する単語群（１つ、または複数の単語）を順次抽出し、検索対象複合語候補格納バッファ部２０７に格納する。さらに、検索対象複合語候補バッファ部２０７に格納されている単語群から最大結合数バッファ部２０４に格納されている結合数以下の複合語を抽出し、検索対象抽出複合語格納バッファ部２０８に格納する。

【００２４】検索対象単語出現頻度算出部１０９は、検索対象単語複合語抽出部２０８により抽出された個々の単語や複合語について、検索対象文書中での出現頻度を算出し、これを検索対象文書の単語情報として検索対象単語情報格納バッファ部２０９に格納する。

【００２５】検索対象単語情報書き込み部１１０は、検索対象単語情報格納バッファ部２０９に格納されている検索対象文書の単語情報を外部記憶装置４に格納する。

【００２６】検索キー文書入力部１１１は、入力装置１から入力された検索キー文書の情報を検索キー文書格納バッファ部２１０に格納する。

【００２７】検索キー単語切り出し部１１２は、検索キー文書格納バッファ部２１０に格納されている検索キー文書からの単語切り出しを行う。そして、その検索キー文書から抽出される全ての単語とその品詞を検索キー文書全文分割単語格納バッファ部２１１に格納する。単語の切り出しは形態素解析などにより行、その文書から抽出される単語の品詞情報を「名詞」、「サ変名詞」、「形容詞」などで表現する。　検索キー単語複合語抽出部１１３は、検索キー全文分割単語格納バッファ部２１１に格納されている全ての単語とその品詞情報の中から、選択品詞情報バッファ部２０２に格納されている品詞情報を参照して、該当する単語群（１つ、または複数の単語）を順次抽出し、検索キー複合語候補格納バッファ部２１２に格納する。さらに、検索キー複合語候補格納バッファ部２１２に格納されている単語群から最大結合数バッファ部に格納されている結合数以下の複合語を抽出し、検索キー抽出複合語格納バッファ部２１３に格納する。

【００２８】検索キー単語出現頻度算出部１１４は、検索キー単語複合語抽出部１１３により抽出された個々の単語や複合語について、検索キー文書中での出現頻度を算出し、これを検索キー文書の単語情報として検索対象単語情報格納バッファ部２１４に格納する。

【００２９】検索対象単語情報読み出し部１１５は、外部記憶装置４に格納されている各検索対象文書の単語情報（単語の出現頻度情報）を１文書毎に呼び出し、検索対象単語情報格納バッファ部２０９に格納する。

【００３０】共通単語抽出部１１６は、検索キー単語情報格納バッファ部２１４に格納されている検索キー文書の単語情報と検索対象単語情報格納バッファ部２０９に

格納されている検索対象文書の単語情報とを比較して、一致する単語の種類と出現頻度情報を共通単語情報格納バッファ部２１５に格納する。

【００３１】類似度算出部１１７は、共通単語情報格納バッファ部２１５に格納されている情報に基づき、検索キー文書と検索対象文書との類似度を算出し、その類似度を算出類似度格納バッファ部２１６に格納する。

【００３２】検索結果出力部１１８は、算出類似度格納バッファ部２１６に格納されている各検索対象文書の類似度値を適宜並べ替えて、検索結果出力バッファ部２１７に格納し、さらに検索結果出力バッファ部２１７の内容を表示装置３に出力する。次に、本実施形態の類似文書検索装置の動作を説明する。

【００３３】最初に検索対象文書データベースの作成手順を図３のフローチャートにより説明する。

【００３４】まず、初期化部１０１により全メモリ部の初期化を行い、外部記憶装置４の品詞辞書と不要語辞書の情報をそれぞれ品詞情報バッファ部２０１、不要語情報バッファ部２０３に格納する（ステップ３０１）。品詞情報バッファ部２０１の構成を図９に、不要語情報バッファ部２０３の構成を図１１にそれぞれ示す。

【００３５】続いて抽出対象品詞設定部１０４が起動され、入力装置１を通じてユーザより抽出する単語の品詞選択の入力を受け付けて、図１０に示すように、抽出対象品詞を選択品詞情報バッファ部２０２に格納する。また、同様に最大結合数設定部１０５が起動され、入力装置１を通じてユーザより抽出する複合語の最大結合数値の入力を受け付けて、図１２に示すように、最大結合数バッファ部２０４に格納する（ステップ３０２）。

【００３６】これらの設定が完了すると、検索対象文書読み出し部１０６が外部記憶装置４から複数のテキスト文書を読み出し、検索対象文書格納バッファ２０５に検索対象文書として格納する（ステップ３０３）。具体例として、例えば、図１３に示すような内容のテキスト文書を検索対象文書として格納されたとする。

【００３７】次に、検索対象単語切り出し部１０７が、検索対象文書格納バッファ部２０５に格納されている検索対象文書について、形態素解析などによって単語の切り出しを行い、切り出した単語とその品詞情報を、図１４に示すように、検索対象全文分割単語格納バッファ部２０６に格納する（ステップ３０４）。

【００３８】続いて検索対象単語複合語抽出部１０８が起動される。検索対象単語複合語抽出部１０８は、検索対象全文分割単語格納バッファ部２０６に格納されている当該検索対象文書の全単語とその品詞情報、選択品詞情報バッファ２０２を参照し、該当する単語（１個以上）を、図１５に示すように、検索対象複合語候補格納バッファ部２０７に格納する。

【００３９】さらに、検索対象単語複合語抽出部１０８は、検索対象複合語候補格納バッファ部２０７に格納されている１個、または複数の単語群、不要語情報バッファ部２０３の不要語情報、そして最大結合数バッファ部２０４の最大結合数値を参照し、１以上かつ最大結合数以下の単語の結合からなる複合語を検索対象文書の中から抽出し、図１６に示すように、検索対象抽出複合語格納バッファ部２０８に全て格納する（ステップ３０５）。なお、ここで抽出される複合語には、単独の単語、つまり結合数１の単語も含むものとする。

【００４０】ここで、検索対象単語複合語抽出部１０８により、単語、複合語が少なくとも１個以上抽出された場合（ステップ３０６）、検索対象単語出現頻度算出部１０９が起動される。検索対象単語出現頻度算出部１０９は、検索対象抽出複合語格納バッファ部２０８に格納されている複合語について、当該検索対象文書中での出現頻度を複合語別に累積し、図１７に示すように、検索対象単語情報格納バッファ部２０９に順次格納する（ステップ３０７）。検索対象語情報格納バッファ部２０９において、複合語（単語も含む）と頻度とは対応して登録されており、例えば、単語「住所録」は当該文書中に４回出現していることを表す。

【００４１】以上の複合語抽出処理と抽出単語出現頻度算出処理は当該文書の文末まで行われる（ステップ３０８）。

【００４２】当該文書の複合語抽出処理と抽出単語出現頻度算出処理が終了すると、図１７に示す検索対象単語情報格納バッファ部２０９に格納された情報は、検索対象文書のデータベースとして外部記憶装置４に蓄積される（テップ３０９）。

【００４３】これで１検索対象文書のデータベースへの蓄積が終了するが、検索対象文書格納バッファ部２０５にまだ検索対象文書が残っている場合、ステップ３０４にもどって、前記同様の文書データベース生成が行われる。検索対象文書が残っていない場合、データベースの生成は終了する。

【００４４】ここで、検索対象単語複合語抽出部１０８による単語複合語抽出（ステップ３０５）の手順を図５、図６、図１４～図１６を使って詳しく説明する。

【００４５】まず、初期化として現結合数に０を代入する（ステップ５１０）。複合語候補となる結合単語を抽出するため、図１４に示す検索対象全文分割単語格納バッファ２０６に記憶されている単語に対応する品詞情報と、図１０に示す選択品詞情報バッファ部２０２にある品詞情報とを比較し（ステップ５０２）、検索対象全文分割単語格納バッファ２０６に記憶されている単語が対象品詞の単語であった場合、図１５に示すように、当該単語を検索対象複合語候補格納バッファ部２０７に格納し（ステップ５０３）、現結合数に１を加える（ステップ５０４）。対象品詞の単語をすべて抽出したらステップ５０５に移る。

【００４６】ステップ５０５では現結合数を調べ、現結

合数が０より大きかった場合は処理を続行し、０であった場合は複合語抽出処理を終了する。

【００４７】現結合数が０より大きい場合は、続いて、複合語抽出のための先頭カウンタと結合数カウンタにそれぞれ１をセットして初期化を行う（ステップ５０６）。

【００４８】ここから、複合語の抽出が、結合数カウンタが最大結合数バッファ部２０４が示す値になるまで以下のように行われる（ステップ５０７）。

【００４９】先頭カウンタが示す検索対象複合語候補格納バッファ部２０７の単語から結合数カウンタの示す単語数の複合語を抽出できる場合（ステップ５０８）、ステップ５０９からステップ５１２にかけて不要語チェックを行う。不要語チェックは、当該複合語を構成する単語と図１１に示す不要語情報バッファ部２０３の単語とを全て比較し、不要語に該当するものがあった場合、その不要語を含む複合語を抽出の対象としない処理を行う（ステップ５１３）。

【００５０】なお、ここでは複合語の一要素となる単語が不要語であった場合、複合語抽出の対象としない処理を行ったが、そうした不要語が複合語の頭に接頭する、あるいは、末尾に接尾する場合にだけ、複合語抽出の対象としない処理を行うようにしてもよい。

【００５１】不要語にあたる単語が、抽出された複合語に含まれない場合は、その複合語を検索対象抽出複合語格納バッファ部２０８に格納する（ステップ５１４）。

【００５２】このときの結合数カウンタが示す単語数の複合語を図１５に示す検索対象複合語候補格納バッファ２０７からすべて抽出する（ステップ５１４）。すべて抽出したら、結合数カウンタに１を加え（ステップ５１５）、先頭から新たな結合数カウンタが示す結合数の複合語の抽出を行う（ステップ５１６）。

【００５３】そして結合数カウンタが最大結合数バッファ部２０４が示す値を超えた場合、複合語抽出を終了する（ステップ５０７）。

【００５４】次に、類似文書の検索手順を、図４のフローチャートにより説明する。

【００５５】まず、初期化部１０１により全メモリ部を初期化し、外部記憶装置４の品詞辞書と不要語辞書の情報をそれぞれ品詞情報バッファ部２０１、不要語情報バッファ部２０３に格納する。（ステップ４０１）。続いて抽出対象品詞設定部１０４が起動され、入力装置１を通じてユーザより抽出する単語の品詞選択の入力を受け付けて抽出対象品詞を選択品詞情報バッファ部２０２に格納する。また、同様に最大結合数設定部１０５が起動され、入力装置１を通じてユーザより抽出する複合語の最大結合数値の入力を受け付けて最大結合数バッファ部２０４に格納する（ステップ４０２）。

【００５６】続いて、検索キー文書入力部１１１が起動され、入力装置１を通じてユーザより検索キーとなる文書の入力を受け付けて検索キー文書格納バッファ部２１０に格納する（ステップ４０３）。具体例として、例えば、図１８に示すような内容のテキスト文書を検索キー文書として格納したとする。

【００５７】次に、検索キー単語切り出し部１１２が、検索キー文書格納バッファ部２１０に格納されている検索キー文書について、形態素解析などによって単語の切り出しを行い、切り出した単語とその品詞情報を検索キー全文分割単語格納バッファ部２１１に格納する（ステップ４０４）。

【００５８】そして、検索キー単語複合語抽出部１１３が起動される。検索キー単語複合語抽出部１１３は、検索キー全文分割単語格納バッファ部２１１に格納されている当該検索キー文書の全単語とその品詞情報、選択品詞情報バッファ２０１，２０２を参照し、該当する単語（１個以上）を検索キー複合語候補格納バッファ部２１２に格納する。さらに、検索キー単語複合語抽出部１１３は、検索キー複合語候補格納バッファ部２１２に格納されている１個、または複数の単語群、不要語情報バッファ部２０３の不要語情報、そして最大結合数バッファ２０４の最大結合数値を参照し、１以上かつ最大結合数以下の単語が結合した複合語を検索キー文書の中から抽出し、これらを検索キー抽出複合語格納バッファ部２１３に全て格納する（ステップ４０５）。

【００５９】なお、ここで抽出される複合語には、単独の単語、つまり結合数１の単語も含むものとする。

【００６０】ここで、検索キー単語複合語抽出部１１３により、単語、複合語が少なくとも１個以上抽出された場合（ステップ４０６）、検索キー単語出現頻度算出部１１４が起動される。検索キー単語出現頻度算出部１１４は、検索キー抽出複合語格納バッファ部２１３に格納されている複合語について、当該検索キー文書中での出現頻度を複合語別に累積し、検索キー単語情報格納バッファ部２１４に順次格納する（ステップ４０７）。図２２に検索キー単語情報格納バッファ部２１４の格納例を示す。この検索キー単語情報格納バッファ部２１４において、複合語（単語も含む）と頻度は対応しており、例えば、単語「筆文字」は当該文書中に３回出現していることを表す。

【００６１】この複合語抽出処理と抽出単語出現頻度算出処理を当該文書の文末まで行う（ステップ４０８）。これで検索キー文書の単語情報の生成が終了する。

【００６２】次に、検索対象単語情報読み出し部１１５が、外部記憶装置４に格納されている各検索対象文書の単語情報を１文書毎に読み込み、検索対象単語情報格納バッファ２０９に格納する（ステップ４０９）。

【００６３】続いて、共通単語抽出部１１６が起動され、検索対象単語情報格納バッファ部２０６と検索キー単語情報格納バッファ部２１４とに共通して格納されている単語、複合語を共通単語情報格納バッファ部２１５

に格納する。具体例を図２３に示す。図１７の検索対象単語情報格納バッファ部２０９と図２２の検索キー単語情報格納バッファ部２１４に共通する単語（複合語）として、「宛名印刷」が抽出され、この「宛名印刷」とその頻度が検索キー側、検索対象側それぞれ１、２というように対応づけて格納する（ステップ４１０）。

【００６４】次に、類似度算出部１１７が、共通単語情報格納バッファ部２１５に格納されている頻度情報に基づき検索キーと検索対象文書との類似度を空間ベクトル法などにより算出し、その類似度値を算出類似度格納バッファ部２１６に格納する（ステップ４１１）。例えば、図２４に示すように、各検索対象文書ごとの類似度が算出類似度格納バッファ部２１６に格納される。

【００６５】全ての検索対象文書について類似度計算が終了すると（ステップ４１２）、検索結果出力部１１８は、算出類似度格納バッファ部２１６に格納されている各検索対象文書ごとの類似度を類似度が高い順に並べ替えて検索結果出力バッファ部２１７に格納し、そのバッファの内容を表示装置３に出力する。出力結果は、例えば、図２６が示すような形で出力される（ステップ４１３）。なお、図２６では類似度値に閾値を設けて表示しているが、類似度値に一定の閾値を設けて、検索結果として表示する検索対象文書の量を制限できるようにしてもよい。

【００６６】これで１検索キー文書の類似文書検索は終了するが、新たに検索キー文書がある場合、ステップ４０２に戻って、同様な処理を行う。検索キー文書がなければ検索処理はこれで終了する（ステップ４１４）。

【００６７】なお、この類似文書検索処理における単語複合語抽出の手順は、検索対象文書データベース作成における処理と対象となる文書による処理部、バッファ部の違いはあるが全く同様の処理である。

【００６８】以上のように、本実施形態の類似文書検索装置では、検索キー文書あるいは検索対象文書に、例えば「筆文字宛名印刷機能」といった複合語が存在するならば、最大結合数を３としたとき「筆文字宛名」「文字宛名印刷」「宛名印刷機能」「筆文字」「文字宛名」「宛名印刷」「印刷機能」などの各所において部分的に連続した複合語が抽出される。このように、特定の内容の文書を特徴付ける異なる複合語を漏れなく抽出することができるので、文書間のより妥当な類似度を計算でき、ユーザの意図する精度の高い類似文書検索を行うことができる。また、検索対象となる文書全体から抽出される単語種の総数が少なくなり、データベースの規模を縮小することができる。

【００６９】さらに、本実施形態の類似文書検索装置では、文書中から複合語を抽出するとき、ユーザにより指定された不要語の単語を含む複合語を無効なものとすることで、ユーザの意図する類似文書検索をさらに高精度に行うことが可能になる。

【００７０】なお、以上説明した類似文書検索装置は，例えば、汎用的なハードウェア環境に、フロッピーディスク、ＣＤ－ＲＯＭなどの記録媒体に記録されたアプリケーションプログラムを追加することによっても提供することが可能である。

【００７１】

【発明の効果】以上説明したように、本発明によれば、文書から抽出された単語の結合により構成される複合語のうち、指定された上限値以下の数の単語により構成されるすべての複合語を検索キー文書および検索対象文書から抽出することで、文書を特徴付ける複合語を漏れなく抽出することができ、検索キー文書と検索対象文書との類似度をより高精度に算出することができる。また、予め指定された不要語の単語を含む複合語を無効なものとすることで、ユーザの意図をさらに反映したより一層高精度な類似文書検索を行うことが可能になる。

【図面の簡単な説明】

【図１】本発明の実施形態である類似文書検索装置のハードウェア構成を示すブロック図

【図２】図１の制御装置の内部構成を示すブロック図

【図３】本実施形態の類似文書検索装置の検索対象文書データベース生成の動作手順を示すフローチャート

【図４】本実施形態の類似文書検索装置の類似文書検索の動作手順を示すフローチャート

【図５】複合語抽出処理の詳細な動作手順を示すフローチャート

【図６】図５と同じく複合語抽出処理の詳細な動作手順を示すフローチャート

【図７】品詞辞書の内容を示す図

【図８】不要語辞書の内容を示す図

【図９】品詞情報バッファの内容を示す図

【図１０】選択品詞情報バッファの内容を示す図

【図１１】不要語情報バッファの内容を示す図

【図１２】最大結合数バッファの内容を示す図

【図１３】検索対象文書格納バッファの内容を示す図

【図１４】検索対象全文分割単語格納バッファの内容を示す図

【図１５】検索対象複合語候補格納バッファの内容を示す図

【図１６】検索対象抽出複合語格納バッファの内容を示す図

【図１７】検索対象単語情報格納バッファの内容を示す図

【図１８】検索キー文書格納バッファの内容を示す図

【図１９】検索キー全文分割単語格納バッファの内容を示す図

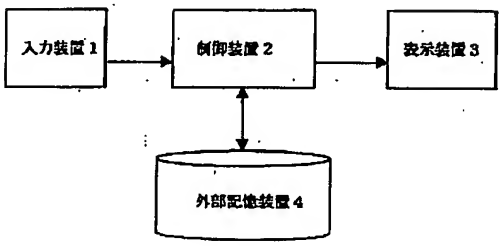
【図２０】検索キー複合語候補格納バッファの内容を示す図

【図２１】検索キー抽出複合語格納バッファの内容を示す図

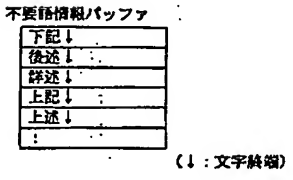
【図２２】検索キー単語情報格納バッファの内容を示す図
【図２３】共通単語情報格納バッファの内容を示す図
【図２４】算出類似度格納バッファの内容を示す図
【図２５】検索結果出力バッファの内容を示す図
【図２６】類似文書検索結果の出力例を示す図
【符号の説明】

１００	制御部
１０１	初期化部
１０２	入力部
１０３	出力部
１０４	抽出対象品詞設定部
１０５	最大結合数設定部
１０６	検索対象文書読み出し部
１０７	検索対象単語切り出し部
１０８	検索対象単語複合語抽出部
１０９	検索対象単語出現頻度算出部
１１０	検索対象単語情報書き込み部
１１１	検索キー文書入力部
１１２	検索キー単語切り出し部
１１３	検索キー単語複合語抽出部
１１４	検索キー単語出現頻度算出部
１１５	検索対象単語情報読み出し部
１１６	共通単語抽出部
１１７	類似度算出部
１１８	検索結果出力部
２００	メモリ部
２０１	品詞情報バッファ部
２０２	選択品詞情報バッファ部
２０３	不要語情報バッファ部
２０４	最大結合数バッファ部
２０５	検索対象文書格納バッファ部
２０６	検索対象全文分割単語格納バッファ部
２０７	検索対象複合語候補格納バッファ部
２０８	検索対象抽出複合語格納バッファ部
２０９	検索対象単語情報格納バッファ部
２１０	検索キー文書格納バッファ部
２１１	検索キー全文分割単語格納バッファ部
２１２	検索キー複合語候補格納バッファ部
２１３	検索キー抽出複合語格納バッファ部
２１４	検索キー単語情報格納バッファ部
２１５	共通単語情報格納バッファ部
２１６	算出類似度格納バッファ部
２１７	検索結果出力バッファ部

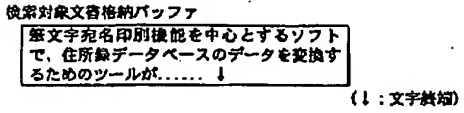
【図１】



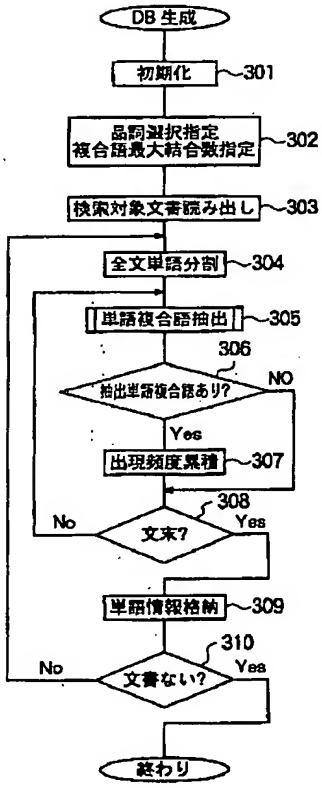
【図１１】



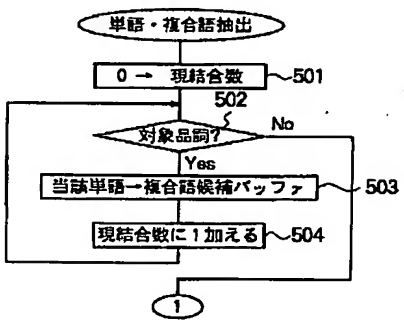
【図１３】



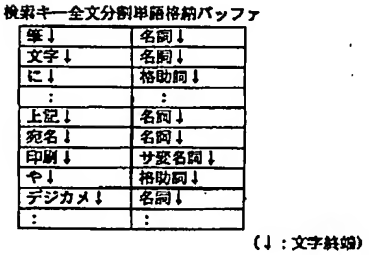
【図３】



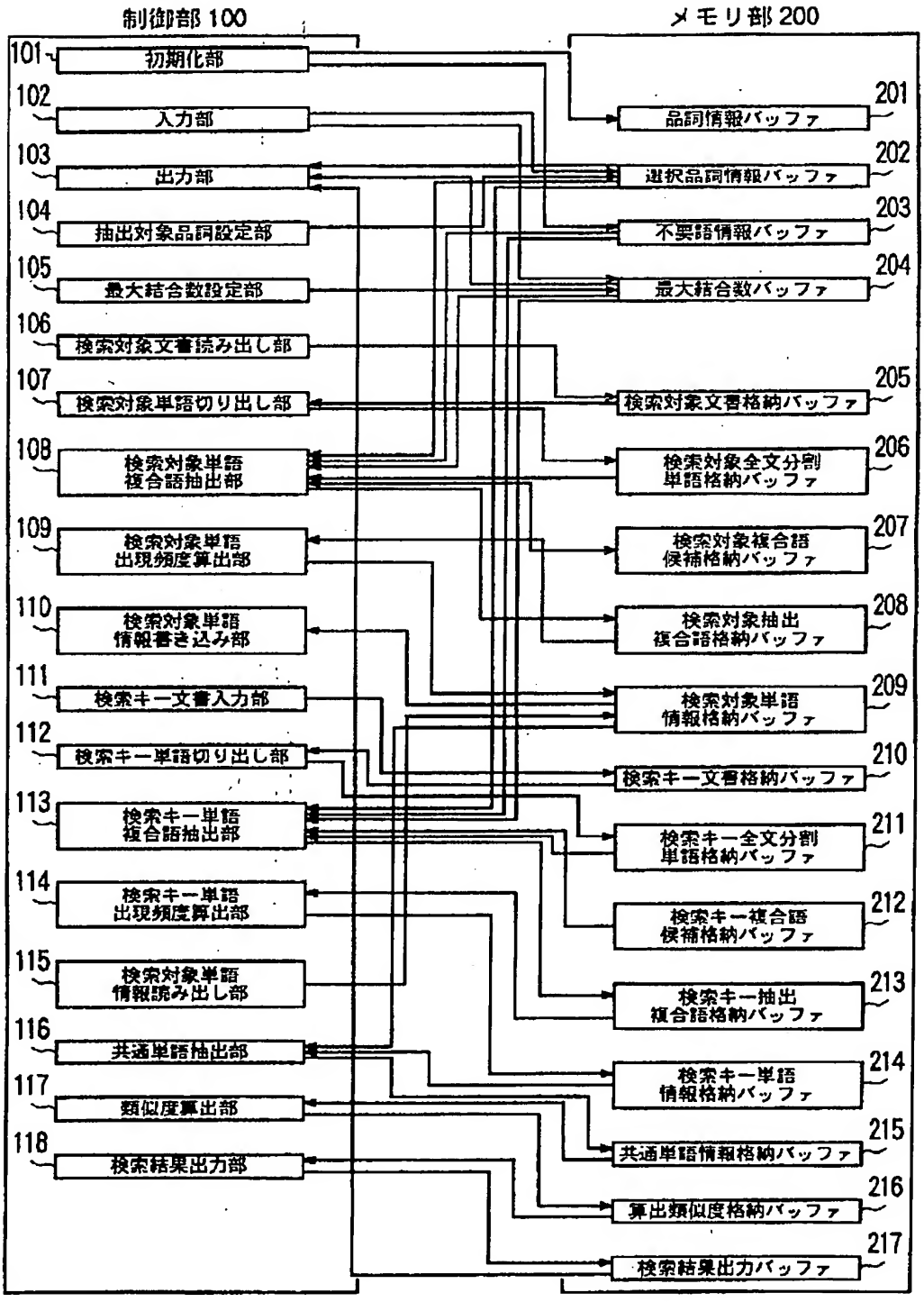
【図５】



【図１９】



【図２】



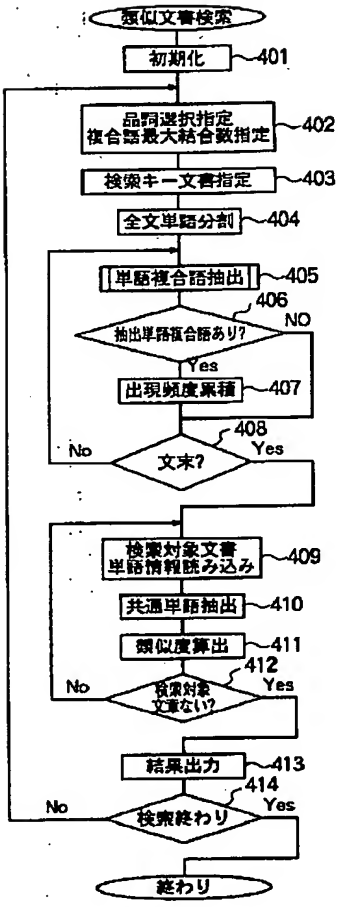
【図１２】

最大結合数バッファ

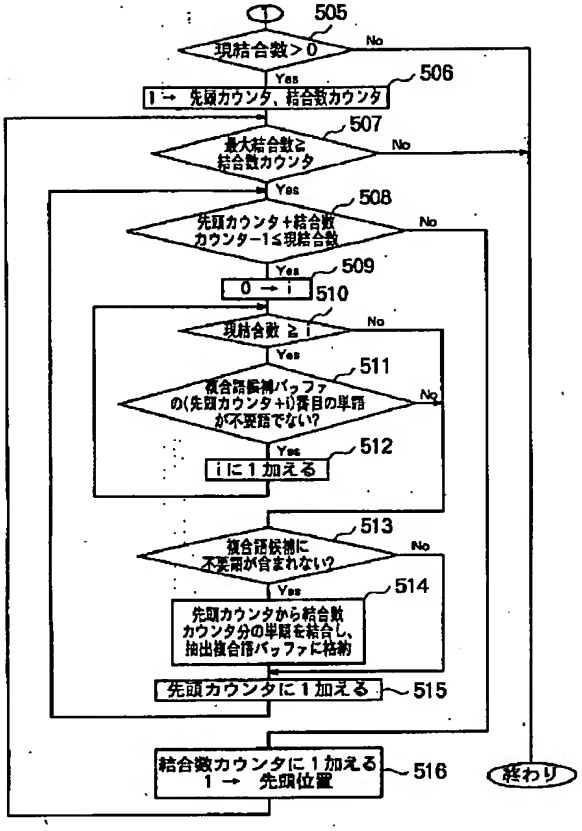
3↓

（↓：文字終端）

【図４】



【図６】



【図７】

【図８】

【図１０】

品詞辞書

品詞
名詞↓
サ変名詞↓
動詞↓
形容詞↓
形容動詞↓
・・

（↓：文字終端）

【図１５】

不要語辞書

登録単語
下記↓
後述↓
詳述↓
上記↓
上述↓
図中↓
前記↓
：：

（↓：文字終端）

選択品詞情報バッファ

名詞↓
サ変名詞↓
形　詞↓
形容動詞↓

（↓：文字終端）

検索対象複合語候補　格バッファ

筆↓	名詞↓
文字↓	名詞↓
宛名↓	名詞↓
印刷↓	サ変名詞↓
機能↓	名詞↓

（↓：文字終端）

【図９】

品詞情報バッファ

名詞↓
サ変名詞↓
形容詞↓
形容動詞↓
：
副詞↓
格助詞↓

（↓：文字終端）

【図１４】

検索対象全文分割単語格納バッファ

筆↓	名詞↓
文字↓	名詞↓
宛名↓	名詞↓
印刷↓	サ変名詞↓
機能↓	名詞↓
も↓	格助詞↓
中心↓	名詞↓
と↓	格助詞↓
：	：

（↓：文字終端）

【図１６】

検索対象抽出複合語格納バッファ

筆文字宛名↓
文字宛名印刷↓
：
宛名印刷機能↓
：
宛名印刷↓
：

（↓：文字終端）

【図１７】

検索対象単語情報格納バッファ

単語	頻度
筆文字宛名↓	2↓
宛名印刷↓	2↓
：	：
住所録↓	4↓
：	：
変換↓	8↓
：	：

（↓：文字終端）

【図１８】

【図２０】

【図２１】

検索キー文書格納バッファ

筆文字フォントによる上記宛名印刷やデジカメ画像からの映像取り込み、住所録のデータベースのインポート／エクスポート機能が充実……↓

（↓：文字終端）

検索キー複合語候補格納バッファ

上記↓	名詞↓
宛名↓	名詞↓
印刷↓	サ変名詞↓

（↓：文字終端）

検索キー抽出複合語格納バッファ

宛名印刷↓
宛名↓
印刷↓

（↓：文字終端）

【図２２】

【図２３】

【図２５】

検索キー単語情報格納バッファ

単語	頻度
筆文字↓	3↓
宛名印刷↓	1↓
：	：
住所録↓	6↓
：	：
機能↓	4↓

共通単語情報格納バッファ

単語	検索キー側頻度	検索対象側頻度
宛名印刷↓	1↓	2↓
住所録↓	6↓	4↓
変換↓	8↓	3↓
機能↓	4↓	1↓
：	：	：

（↓：文字終端）

検索結果出力バッファ

類似検索対象文書
3↓
51↓
289↓
：

（↓：文字終端）

【図２４】

【図２６】

算出類似度格納バッファ

検索対象文書番号	類似度
1↓	0.03198↓
2↓	0.16771↓
3↓	0.49534↓
：	：

（↓：文字終端）

検索結果出力例

類似文書検索結果

<文書番号>
3
51
289
：

フロントページの続き

(72)発明者　仁科　卓哉
東京都青梅市新町1381番地１　東芝コンピュータエンジニアリング株式会社内

(72)発明者　中本　幸夫
東京都青梅市新町1381番地１　東芝コンピュータエンジニアリング株式会社内